TS-EAS EAC-CPF Berlin meeting

Attendees (in person)

Name	Institution	Role
Caitlin Rizzo	Penn State University, US	EAC-CPF team
Cory Nimer	Brigham Young University, US	Documentation team
Gerhard Müller	Staatsbibliothek zu Berlin, DE	TS-EAS ex officio
Joost van Koutrik	Het Utrechts Archief, NL	TS-EAS
Karin Bredenberg	Kommunalförbundet Sydarkivera, SE	TS-EAS co-chair
Kerstin Arnold	Archives Portal Europe Foundation, EU	EAD team lead
Regine Heberlein	Princeton University, US	former EAC-CPF team member
Silke Jagodzinski	Geheimes Staatsarchiv Preußischer Kulturbesitz, DE	EAC-CPF team lead

Tuesday, 10 March 2020

Topic: Relations

Gerhards starts with an introduction, which was based on his key issue paper [Topic: Relations](#) concerning the revision of encoding relationships within EAC-CPF. His main objectives are:

- simplification and clarification of encoding
- ensuring cross-domain interoperability
- reconciliation with EAD3

Short introduction to SoNAR

A very early visualization of concepts elaborated by the project [SoNAR (IDH)](#) that has been funded by the German Research Foundation was demonstrated. The visualizations included geographical and occupational distributions of a sample of 3,068 agents linked with archival records encoded with EAD. [Kalliope](#) encodes relations between archival resources and agents that created, received, or are subject by using authority records with shared persistent identifiers. Kalliope reuses authority records by transforming data from MARC21 to EAC-CPF.

Authority records about an agent may include explicit relations between the agent and

    A. persons (e.g. family),
    B. corporate bodies (e.g. affiliations),
    C. conferences,
    D. time (e.g. birth date),
    E. occupation (e.g. a profession a person is identified with), and
    F. places (e.g. headquarter of a corporate body)

these related entity types are, whenever possible, identified by persistent identifiers of shared authority records. The Integrated Authority File of the [German National Library has an ontology](#) available in OWL online to support linking and sharing data.

EAD in turn, has explicit relations between a record resource and i.a. an agent encoded ([<controlaccess>](#) sub-elements with @role, @source, @authfilenumber, @normal). Information about a repository is also recorded in sub-elements of the [<repository>](#) field using @role, @encodinganalog, @source, and @authfilenumber (for ISIL-Code). @encodinganalog is generally used for an internal database number, while the @authfilenumber is used for authority file numbers, e.g.

```
<persname role="Verfasser" encodinganalog="DE-611-PS-174638"
source="GND" authfilenumber="11851508X" normal="Bretano,
Sophie">Brenato, Sophie (1776-1800)</persname>
```

Relationship types in the context of archival description

EAC-CPF and EAD3 currently support the encoding of data about relations between the following entity types:

EAC-CPF	EAD3
CPF to CPF	Resource to Resource
CPF to Function	Resource to Function
CPF to Resource	Resource to CPF
CPF to Time	Resource to Time
CPF to Place	Resource to Place
CPF to Subject (Mandate, Function, Occupation, Legal Status)	Resource to Subject (Subject, Function, Occupation, Legal Status, Works)
CPF to Local Description	Resource to Works (Monographs/Periodicals) (bibliography)
	Resource to Instance (dao)

- CPF to CPF and Resource to CPF in this context also include agents that are responsible for maintaining records (holding organizations)

Not available are events (e.g. marriage), where the event itself is identified with a Place, a Time, or other related entities. EAD3 has the Chronology Item Set <chronitemset>, which was introduced to allow for bundling several places and several events to the same date field. It includes the element <event>, which, however, encodes unstructured text only.

The scope of modeling events with the Encoded Archival Standards should be addressed at the Chicago meeting.

In order to encode data about relations a set of elements is needed:
- Name of the related entity (required)
- Type of related entity
- Type of relationship
- Role of entity in relationship
    - Characterizes the context of each of two entities in a relation (e.g. if type is hierarchical, one entity is the superior and the other the subordinate)
- Time or time frame of a relationship
- Place of a relationship
- Source of the relationship statement

Relationships currently can be recorded using unstructured text, structured text and URIs. It would be preferable to have a method of differentiating between these approaches so that it can be clear that URIs are available.

Relationships in EAC-CPF descriptions section are generally straightforward by sharing the same child elements and attributes. Only the Place <place> uses different child elements.

When recording relationships, we need to be aware of the depth of information available. Generally these are recorded in simple fashion A → B. Additional information about the relationship type (e.g., not just "associative", but "friend") might be difficult to obtain or uncertain/unconfirmed. In some cases, particularly for events, there may also be an associated date and place for the relationship.

We need to look at the relation section, perhaps decluttering in the descriptive section of EAC-CPF, to increase use of structured relationships section. Whereas the descriptive section could concentrate on simple relationships, ideally with URIs, but also allowing for literals, the relations section could concentrate on describing the relationship itself. Here again the values can be both available as literal only or as resource represented by URI.

It would depend on the level of detail, where a relationship is recorded: when only few data are available it might make more sense to have it in the descriptive section, while in other cases relations can be described in detail in the relations section.

Both quality of data and level of detail should be in tag library examples.

The proposal calls for replacing current relations child elements of <relations> (<cpfRelation>, <resourceRelation>, <functionRelation>) by a unifying relations child element <relation>. A first intention was to reuse following child elements: <relationEntry>, <date>, <dateRange>, <dateSet> and to introduce new child elements: <relationType>, <relationRole>, <place> and <assertion>.

Proposal

To **unify** the relations child element by using a **<relation>** element, similiar or other singular/plural elements was agreed.

By solving the assertion description with another solution, the element for assertion or evidence wouldn't be used. Instead the new attributes to link to a source, **@sourcelink**, and to a maintenance event, **@maintenanceeventlink**, will be used within <relation>. The exact attribute names might be changed.

The element **<relationEntry>** is used to identify the related entity with a text. It was agreed to **rename** the element **to <targetentity>** in order to be precise in the elements name.

An attribute for the type of the target entity must be used with limited values, i.e. **@targettype** with values corporatebody, person, family, resource and function, within <targetentity>. This attribute corresponds to the EAD3 attribute @relationtype.
This approach allows to extend the entity types, but also simplifies the encoding. New entity types might be added by including the Records in Contexts approach in the EAC-CPF schema or by introducing a new entity description for events.

In order to encode the source of the related entity and the value by URIs, 3 **new optional attributes** should be introduced to encode relations:

- **@vocabularysource**: text or literal, where the related entity comes from,
- **@vocabularysourceuri**: URI of the source, and
- **@valueuri**: URI of the related or target entity at the mentioned source.

The attribute names might be changed later on.

The element <targetentity> does not contain text or mixed content, but follows the name entry approach with a mandatory child element **<part>**. The <part> element contains text and is repeatable.

The new optional element **<relationtype>** is used to encode the type of a relation, next to the attribute @targettype within <targetentity>. The element contains text and is repeatable. 3 new optional attributes are used to encode URI of the elements content. The element <relationtype> might contain local terms, terms in different languages or other text to provide a description of the relation type.

A new optional element **<targetrole>** is used to specify the role of the related entity within the relation. It is a repeatable text element with optional attributes to encode the URIs of the elements content.

Next to these elements, the existing optional child elements for dates (**<date>**, **<dateRange>**, **<dateSet>**) will be kept.

The distinct encoding of a **place** connected to the related entity wasn't discussed during the meeting and will be treated during a virtual team meeting.

Other existing elements within the relations part (<descriptiveNote>, <objectBinWrap>, <objectXMLWrap>) in the current schema haven't been part of the new proposal. Their usage will be discussed during a virtual team meeting.

Example encoding:

```
<relations>
      <relation id="1" sourceLink="#43" maintenanceLink="#43">
        <targetEntity targetType="person" vocabularySource="GND"
                   vocabularySourceURI="http://d-nb.info/gnd/"
                            valueURI="http://d-
nb.info/gnd/1190671301">
              <part localType="family">Arendt</part>
              <part localType="given">Paul</part>
              <part localType="dates">1873-1913</part>
```

```
            </targetEntity>
            <relationType vocabularySource="GND"
                    vocabularySourceURI="https://d-
    nb.info/standards/elementset/gnd#"              valueURI="https://d-
    nb.info/standards/elementset/gnd#Family"
            localType="text" lang="en">Family</relationType>
            <relationType vocabularySource="GND"
                    vocabularySourceURI="https://d-
    nb.info/standards/elementset/gnd#"              valueURI="https://d-
    nb.info/standards/elementset/gnd#Family"
            localType="code">bezf</relationType>
            <targetRole vocabularySource="schema.org"
                    vocabularySourceURI="http://schema.org"
                        valueURI="http://schema.org/parent"
    localType="text"
            lang="en">parent</targetRole">
            <targetRole vocabularySource="schema.org"
                    vocabularySourceURI="http://schema.org"
                        valueURI="http://schema.org/parent"

            localType="code">par</targetRole>
                <date>...</date>
                <place>...</place>
                <descriptiveNote>...</descriptiveNote>
            </relation>
</relations>
```

[Additional examples of resource, etc. needed to complete proposal.]

Decision:

1. Proposal as described above will be distributed with the call for comments on the draft revision, rather than being sent out to the larger committee beforehand.
2. Rename <relationEntry> to <targetEntity>.
3. Add element <relationType> to encode/describe the type of the relation.
4. Add element <targetRole> to encode the role of the related entity.
5. Keep date and place information, also descriptive note element.

Topic: Referencing

The topic of referencing attributes was postponed and nearly solved by decisions to use @...link attributes for assertion description and relations. A paper on this topic summarises the status.

Remaining open questions:

1. How to name the attributes: @...link or @...ref?
2. Where to add and which attributes exactly?
    a. Would we envisage having a general @target attribute as in EAD3? Or mainly context-specific targeting attributes?

RiC-CM and RiC-O

Online presentation on RiC-CM and RiC-O provided by Florence.

The slides are from the conference held at BNF in January, translated to English. The slides are available under CC BY 4.0 and can be reused with this license.

Introduction

ICA had published the 'old' standards from 1994 to 2008. There were some inconsistencies and overlaps, like the "Administrative/biographical history" element in ISAD(G) and the "History" element in ISAAR(CPF). Other problems: relations as defined in ISAAR(CPF) are not yet present in ISAD(G).

ICA decided to develop a new global standard, Records in Contexts (RiC), which had to be backwards compatible with the existing four standards. The differences are:

RiC focusses on entities and on defining them, where ISAD(G) etc. focus on description rules. The glossary in RiC is significantly smaller. In the current ICA standards glossaries (thus definitions of concepts) are present but as, partly quite short, introductory sections. RiC-CM is mainly about defining each of the entities that exist in the world of archives.

These entities are (re)defined, added properties (attributes) and relations between them. Relations are core to the model of RiC, as opposed to the description being core to ISAD(G) etc.

The four previous standards were too few-dimensional to accommodate the multiple layers of contexts (see slide 11). Defining more distinct entities in RiC as well as their relations, would make them 'easier', but also potentially more nuanced and flexible.

See the example on slide 13. This graph shows the relations focused on the president of Radio France. In the finding aid some series were described about Radio France, the mandates, functions and others, grouped together as entities and their relations. Supposedly this form might be more usable to the general public than the 'traditional' finding aids. Note that these entities have relations, *pointing to* other entities. Any of these entities can provide an access point to enter the information system.

See slide 16 about the development of RiC by EGAD. The members 'all over the world' means the Americas, Europe and Africa. One of the main differences from the previous EGAD group is that this group has to continuously maintain and develop the standard going forward.

RiC is currently comprised of three parts: Conceptual Model (CM), Ontology (O, technical specification of the CM), Application guidelines. RiC allows to see the archival world as an

oriented graph, because it consists of entities and relations. EGAD hopes that RiC-O will not be the only technical representation of RiC-CM. There will undoubtedly be other representations. According to Florence, EAD and EAC-CPF should be made compliant with RiC-CM, as RiC-CM replaces ISAD(G) and ISAAR(CPF), on which EAD and EAC-CPF are based.

EAD and EAC-CPF could therefore be viewed as a subset of the Ontology. Before RiC-O there was no internationally supported domain ontology for archives. (There were a number of local initiatives.)

Work on the Application Guidelines is still to be started. This could encompass training courses, examples, etc.

See slide 19 for the development history and roadmap. On the first version of RiC-CM, EGAD received a lot of comments. These were processed in version 0.2 of RiC-CM. Version 0.2 is to be regarded as the only leading version as of its publication.

RiC-O can now also be found on Github. Florence will keep TS-EAS updated on news in this regard. People can create questions and issues on the GitHub site.

EGAD is also preparing the full public release of RiC-CM and RiC-O version 1.0 in November at the ICA conference in Abu Dhabi. Current 0.2 CM and 0.1 O are presumably very similar to what versions 1.0 will be.

Ric-CM 0.2 has 22 entities, grouped in four hierarchies. Main difference with ISAAR(CPF): Apart from CPF, also Positions (E12): functional role of a person in a group, e.g. the president of a company; and Mechanisms (E13), e.g. systems or algorithms.

Main difference with ISDF: Activity (E15) as a broad definition, and a subcategory of an Event (E14), covering all processes, activities, functions, etc.

For Record Resources (E2), there is a distinction between individual Records (E4) and aggregations defined as Record Sets (E3). These entities have a total of 41 attributes. Some of these are repeatable or extensible. See the RiC-CM 0.2 public release document for a specification of each attribute.

There are 78 relations (and the inverse ones) in a poly-hierarchical system, see slide 29. These also have attributes, as detailed in the RiC-CM document. See representation in the chart on slides 32 and further. Here, the relations are represented two-directionally. In the next-to-last and last graphs, the relations are represented as one-dimensional.

From RiC-CM to RiC-O

See also slide 44 and on. RiC-O is a domain ontology for archives, compassing what archivists would use for describing and maintaining archives. It should be easy to maintain and be mappable with other standards. It should be immediately usable and convertible from existing metadata sets. Existing finding aids in EAD and EAC-CPF have been converted to RiC-O. Example given on slide 48 is a query of corporate bodies existing in the nineteenth century in France.

From theory to practise

The URI for the Ontology on the ICA EGAD webpage leads to a version that is readable for human readers (https://www.ica.org/standards/RiC/RiC-O_v0-1.html). The OWL-ontology is also available. There are more classes in RiC-O than there are entities and attributes in RiC-CM (see slides 55 and 57). Example given in the presentation is the relation by class, between the president of Radio France (person and position) with Radio France as a Corporate Body.

Several events have been planned in the next few months in France.

In closing, Florence suggests that EAC-CPF can be updated to be more compliant with RiC. Implementing RiC is reliant on that. In this respect, it is not about re-inventing the wheel, but keeping it moving and improving on it.

She has submitted a document with a number of ideas connected to the work being done on EAC-CPF.

Ontology and schema: usage, differences and approach: Karin Bredenberg

PREMIS Ontology and Shema

This is the international de facto standard for digital preservation, maintained by a group of experts from Europe and the US. Karin is the PREMIS chair.

The Object (see presentation) is of 4 different "types": intellectual entity, file, bitstream and representation. The intellectual entity can be described in PREMIS as of version 3. Relations can be defined two-way, with the exclusion of Agent to Object. One of the presumptions is that data that is being maintained in other systems, should not be maintained in the PREMIS system to make it least cumbersome as possible.

PREMIS does not prescribe its implementation, but when it comes to transfer there is an XML Schema and an RDF in place to adhere to. Usage of RDF presupposes usage of a triple store. Other expressions are also acceptable, depending on which works best for the institution that's using it. Central questions here: how is your system built, and where are you sending the information? (For different incarnations of the given example, see presentation.)

Next actions for the PREMIS agenda: Investigate mapping the RiC-O record with the PREMIS object. Seen from a RiC-perspective, there is a good match; but the PREMIS ontology was not used for RiC-O.

Question: does Karin have any experience with usage of PREMIS ontology by others?

Someone from Princeton was involved with the ontology. In most cases, the XML-expression

For BNF: they have a huge triple store, but other institutions typically do not.

Discussion (impression)

There is some hesitation behind remodelling the EAS to the current version of RiC, mostly because the current model is still a 'moving target' and some of the elements (notably Events/Activities) are considerably different from the previous standards and previous

version of RiC-CM. It is however the view of EGAD that the existing schemas based on ISAD and ISAAR should in future be compliant with RiC-CM.

RiC-CM will be the new guideline from EGAD for the archival community. EGAD provides the community with RiC-O an OWL Ontology for implementing it. On the other hand, there are the current EAS-standards.

It could be crucial to align the current EAD and EAC-CPF to RiC-CM. This is not a question of whether we should, but more one of timing. What does 'align' mean? One of the strengths of RiC is that it is object oriented. The remodelling of EAS to be less description oriented to more object and relation oriented in itself, makes it more aligned to the modular approach as defined in RiC-CM. These standards need not compete. Where is RiC-O positioned, as an instantiation of RiC-CM; and the EAS will become another? Or as the only instantiation?

The current approach as discussed this morning moves to the object approach as espoused by RiC-CM. Current problem is that it is still under development, and the current planning for the revision of EAC-CPF is by late 2020. The approach as of now is to make EAS more modular.

One of the other issues is that RiC-CM and RiC-O are developed and will presumably be maintained with ICA funding. The EAS is mostly self-funded by its members. It is even difficult for the EAS team to get answers from ICA to its questions. This is an issue that needs to be resolved between SAA and ICA.

TS-EAS: schema development and shared schema

Mark explained the current development of the shared schema. The current idea is that we will move forward with a single GitHub repository that will be able to generate EAC-CPF and EAD3 as well as any additional EAS in future. With this approach, we will still be able to retain differences between the schema deliverables (e.g. filedesc in EAD3 but not EAC), if desired and when required, but we will be concentrate on specifying what the differences are up front, rather than being forced to constantly review and attempt to align any drift in definitions between the different driver schemas, which is far more complex as the current process shows.

A very important decision that we still need to make: do we want to retain separate namespaces for EAD3 2.0 and EAC-CPF 2.0, or do we want a single namespace for the EAS standards. There are advantages and disadvantages to both options. The biggest advantage is that we wouldn't need to create new namespaces for any additional EAS schemas (but if there are no others, then there is no advantage). The biggest disadvantage is that we would need to come up with yet another namespace if we go with the single EAS namespace.

A number of issues are relevant here:

Schema principles are still in draft and being worked on. These have been broken down in three main aspects:

- Schema design / technical principles;
- Tag library principles for examples and definitions;
- General standard definition principles, e.g. same name means the same model and does the same thing, don't use the same name for elements and attributes.

What is the schedule for these principles? A number of principles on schema design need to be written down and agreed upon. We do have various points in this meeting where we talk about conventions on a higher level. We should be able to have a system of quality control on decisions on EAS. A set of principles should make this control possible and easier to communicate more 'uniformly'.

Spelling question

In EAC-CPF, CamelCasing is used to make it more userFriendly. Suggested is use of lower case going forward. There should be no ambiguity in terms when only lower case is used; so avoid creating new terms with other meaning when combining two separate terms in one element or attribute name.

Decision: The team agrees on the use of lower case.

XML namespace

First approach was to remove name spaces, as has been done in EAD3. What was the argument for removing it?
This was coming from the XLink perspective. The reason for removal in EAD3 was to not be dependent on anything. The elements that we're using in EAC-CPF are conventions anyway.

Discussion focussed on @xml:id and @xml:lang. @xml:lang is still being used in EAC-CPF, but not anymore in EAD3. It is an obvious element in any type of XML documents; so it would be easily mapped to for example MODS or METS. It would be easier to remove XLink than XML.

We don't know of any other standards that have removed XML name spaces. If we introduce XML to EAD3, all elements would need to be 'changed'. And we would need to explain why the XML name spaces are being adhered to and not any others. Mark proposes to add XML base and XML name spaces. Kerstin stresses the need to be consequent about the choices being made and consequently carrying them out.

Decision: Hand a decision over to the Schema team to analyse the benefits and consequences of usage or removing of xml ns.

Wrap-up and summary

Daily summary for TS-EAS members in a virtual session.

Decisions

Topic	Decision	Issue
Spelling	The team agrees on the use of lower case for element and attribute names.	

	Change abbreviated names in EAC-CPF into descriptive names.	